Carnegie Mellon University

Heinz College

INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Data Mining Final Project
Stop and Frisk

Team: Olivia Hao, Lohith Chiluka and Camila Garcia
Heinz College of Information Systems and Public Policy
Carnegie Mellon University

Today's Agenda

1. Project Overview
2. EDA Sample
3. Questions Walk-Through
4. Analysis Summary

Project Overview

The New York City stop-and-frisk program is a practice of the New York City Police Department in which a police officer who suspects a person has committed, is committing, or is about to commit a felony or a penal law misdemeanor, stops and questions that person, and, if the officer suspects he or she is in danger of physical injury, frisks the person stopped for weapons.

Over the years, the program has caused controversies related to racial profiling. Claims have been made that African-American and Hispanic individuals were stopped more frequently than whites, while the program failed to reduce robbery, burglary, or other crimes.

Project Overview

The purpose of this project will be to use different models to classify three of our variables in order to predict accurately potential outcomes of stop and frisk for each variable in an unseen dataset. As such we will attempt to answer the following questions:

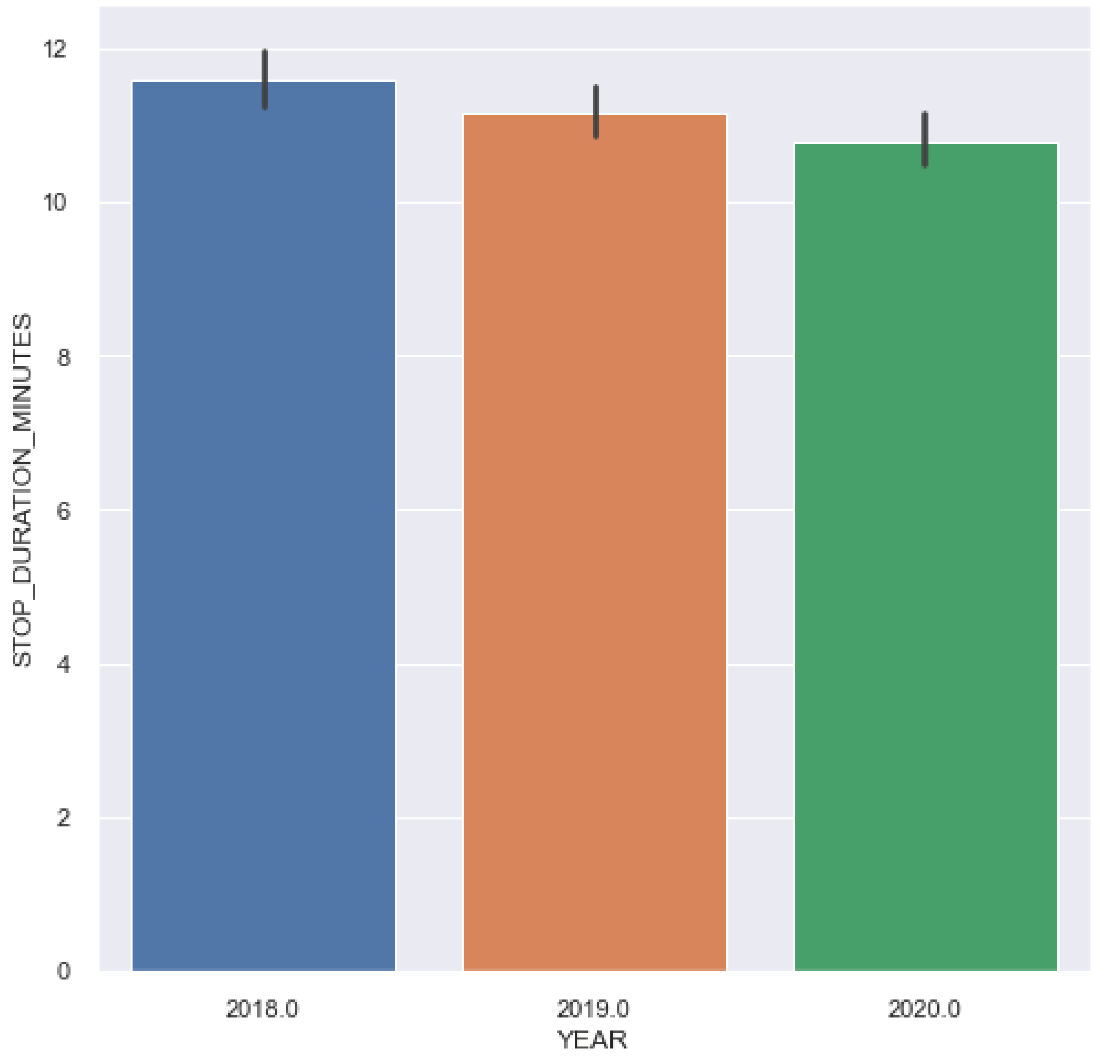
1) Considering the number of individuals stopped in the past 3 years, if an individual were to be stopped, can we predict in which of the following classes they would fall into:

- 0-30 minutes
- 30-60 minutes
- 60+ minutes

2) Accounting for the various factors that could be involved in the reasoning behind an individual being stopped and frisked, how accurately can we predict that an individual who is stopped would fall into the ‘Frisked’ class or the ‘Not-Frisked class?

3) Can we predict that an individual who has been stopped is also searched and what are the features that are most significant to an individual being stopped and searched?

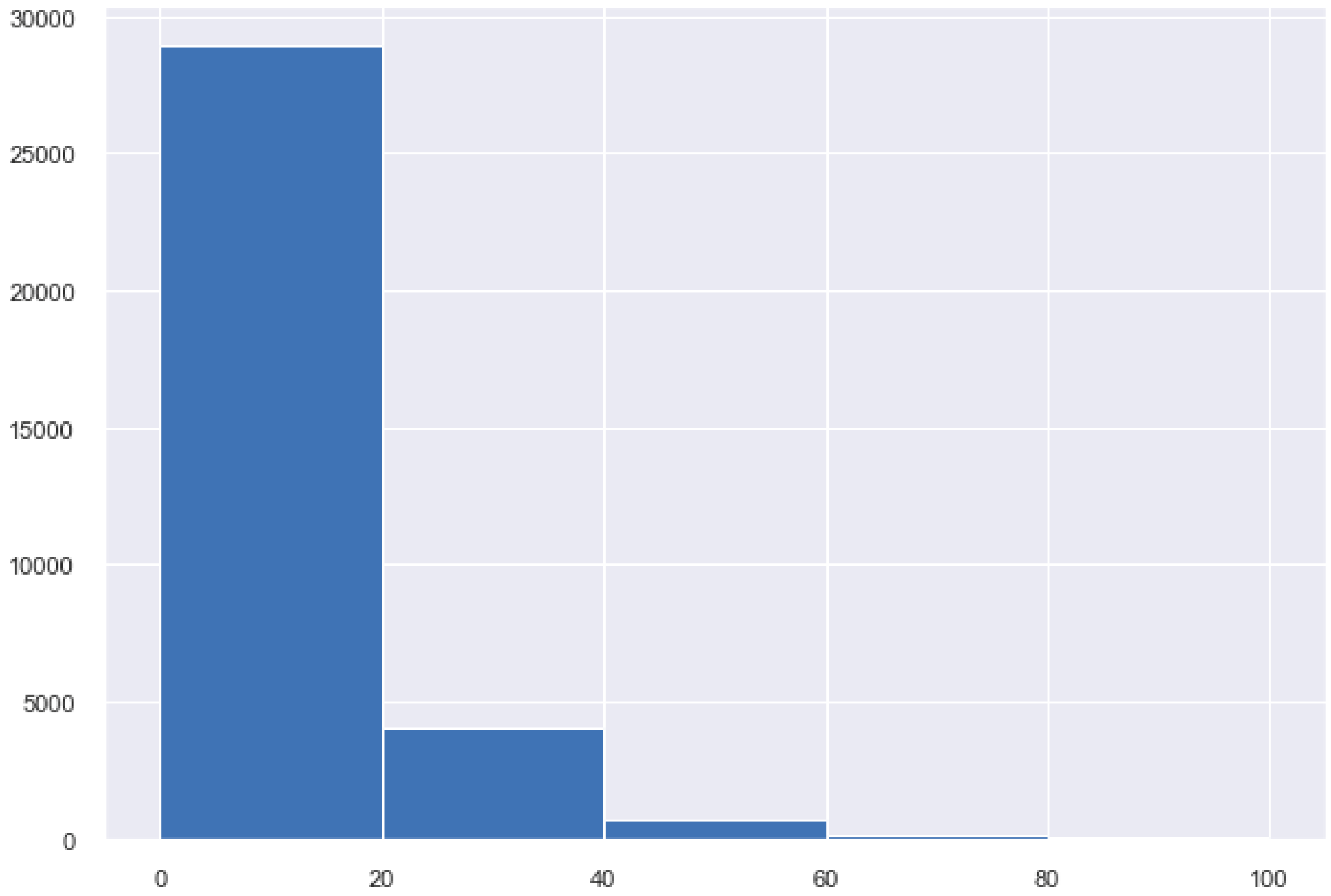
EDA Sample - Average Stop Duration By Year



The average stop duration has seen a declining trend in the recent years. The average stop time for the three years seems to be around ***11 minutes*** *with a decrease of less than a minute in* ***2 years****.*

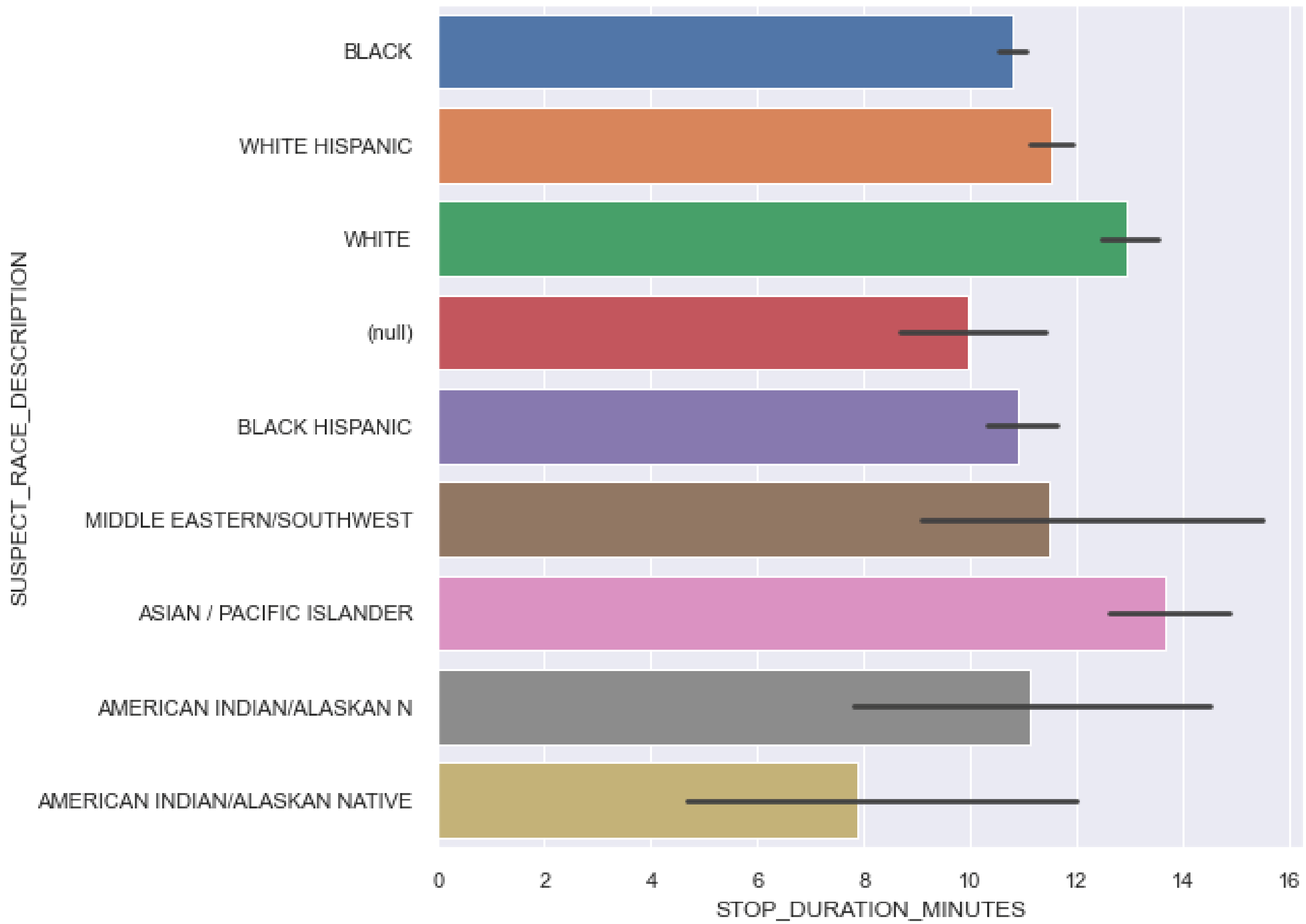
Carnegie Mellon University HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

EDA Sample - Histogram of Stop Duration Distribution



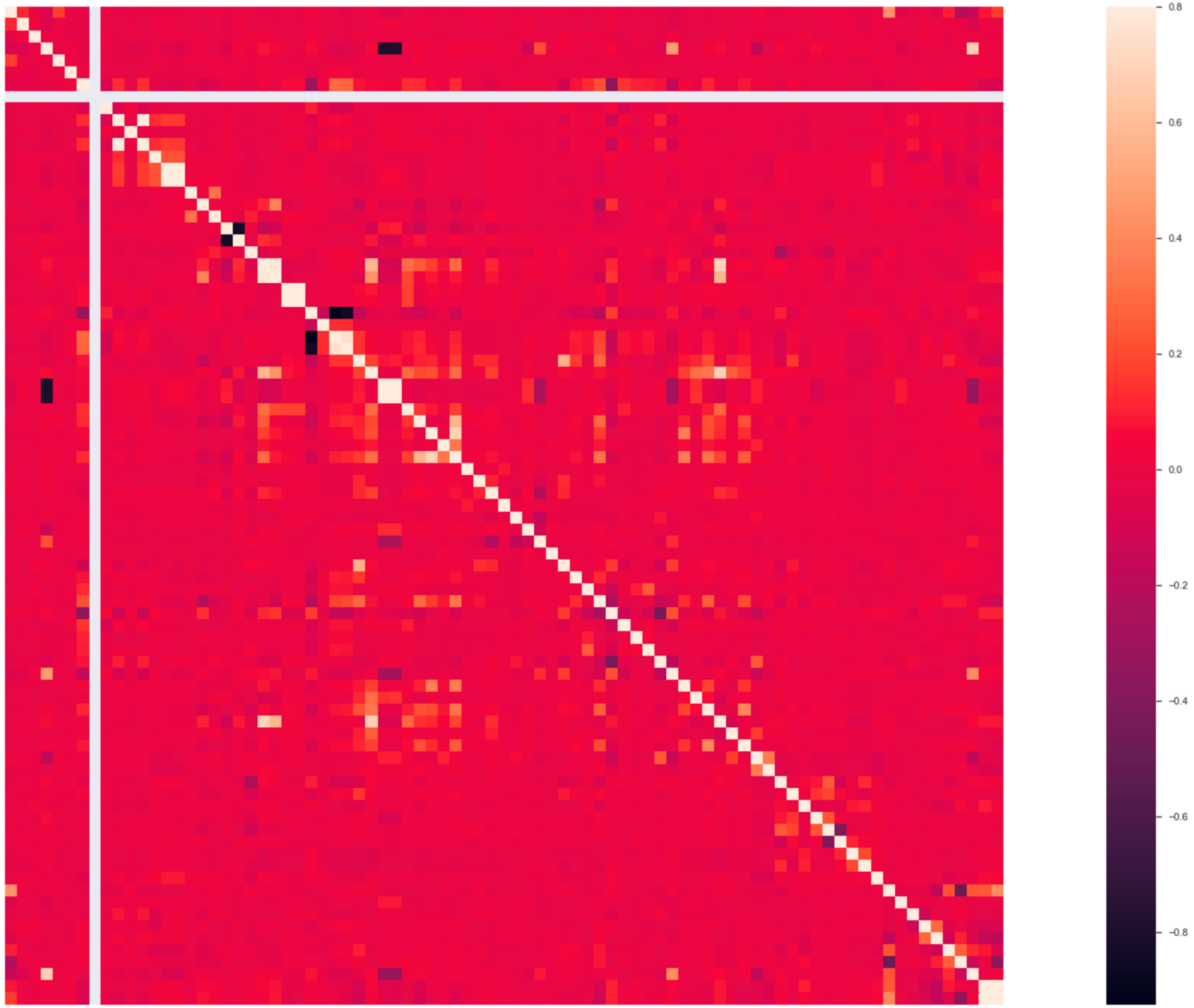
The data on stop duration is ***extremely imbalanced*** *- with the majority of data in the* ***0 to 20 minute range****. This will be an important consideration to make when performing classification for this particular variable.*

EDA Sample - Stop Duration By Race



By assessing the average stop duration broken down by race, we see that asian / pacific islanders are stopped for the longest duration, followed by white and white hispanic. The lowest stop duration seems to be among Native Americans. However, we must acknowledge that there is a ***large chunk of unlabeled record*** *and this breakdown may not be 100% aligned with reality.*

EDA Sample - Correlation Matrix



A preliminary assessment of the correlation matrix indicates that with the exception of a few variables, ***most of the variables are not heavily correlated*** *with each other. As such, we will consider all of our variables for classification models.*

Question 1 - Classification for Stop Duration

Oversampling was applied to these models given the previously mentioned data imbalances. This technique could potentially provide better results into out models.

Summary of models used:

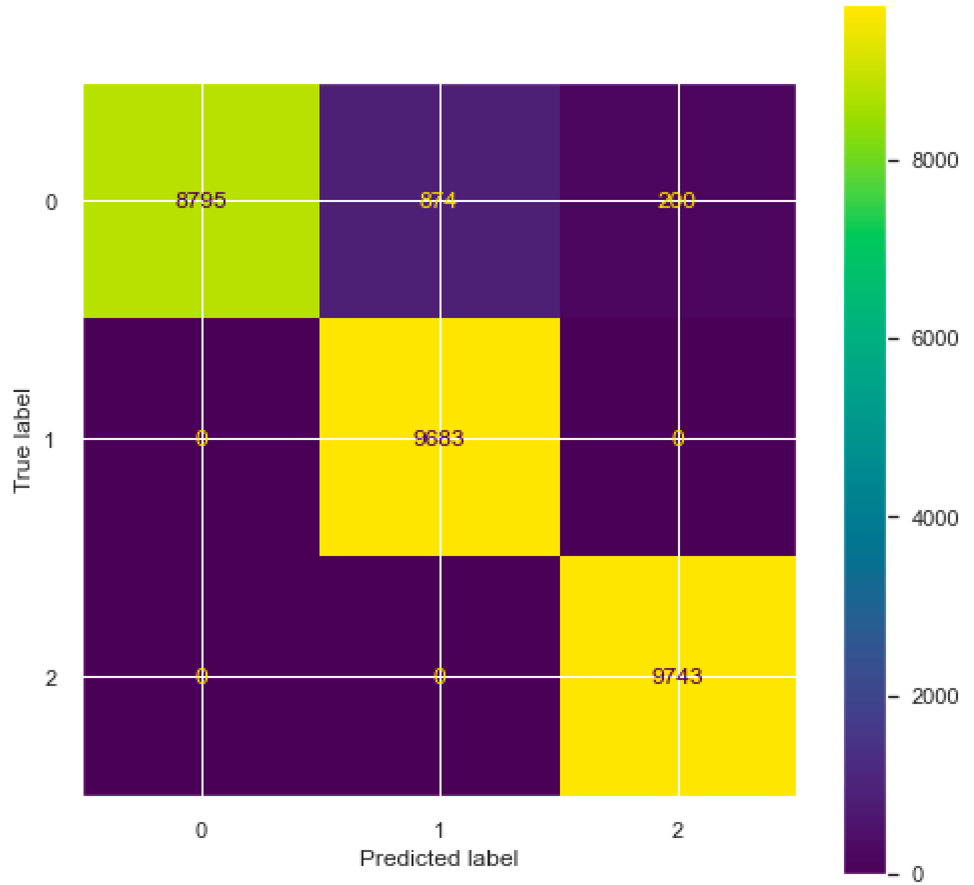
- KNN without oversampling
- KNN with oversampling
- Naive Bayes without oversampling
- Naive Bayes with oversampling
- Logistic Regression without oversampling
- Logistic Regression with oversampling

Model 1 - KNN

After Oversampling

- Once the data has been uniformly distributed throughout the three classes, the results of the knn classification are much better. We can see a similar support score across the classes which is indicative of a evenly distributed dataset, and the resultant scores (accuracy, precision, recall and f1) are all significantly higher that pre-oversampling.
- With oversampling the data, we were able to achieve the best result with KNN.

	precision	recall	f1-score	support
0-30 Min	1.000	0.891	0.942	9869
31-60 Min	0.917	1.000	0.957	9683
60+ Min	0.980	1.000	0.990	9743
accuracy			0.963	29295
macro avg	0.966	0.964	0.963	29295
weighted avg	0.966	0.963	0.963	29295

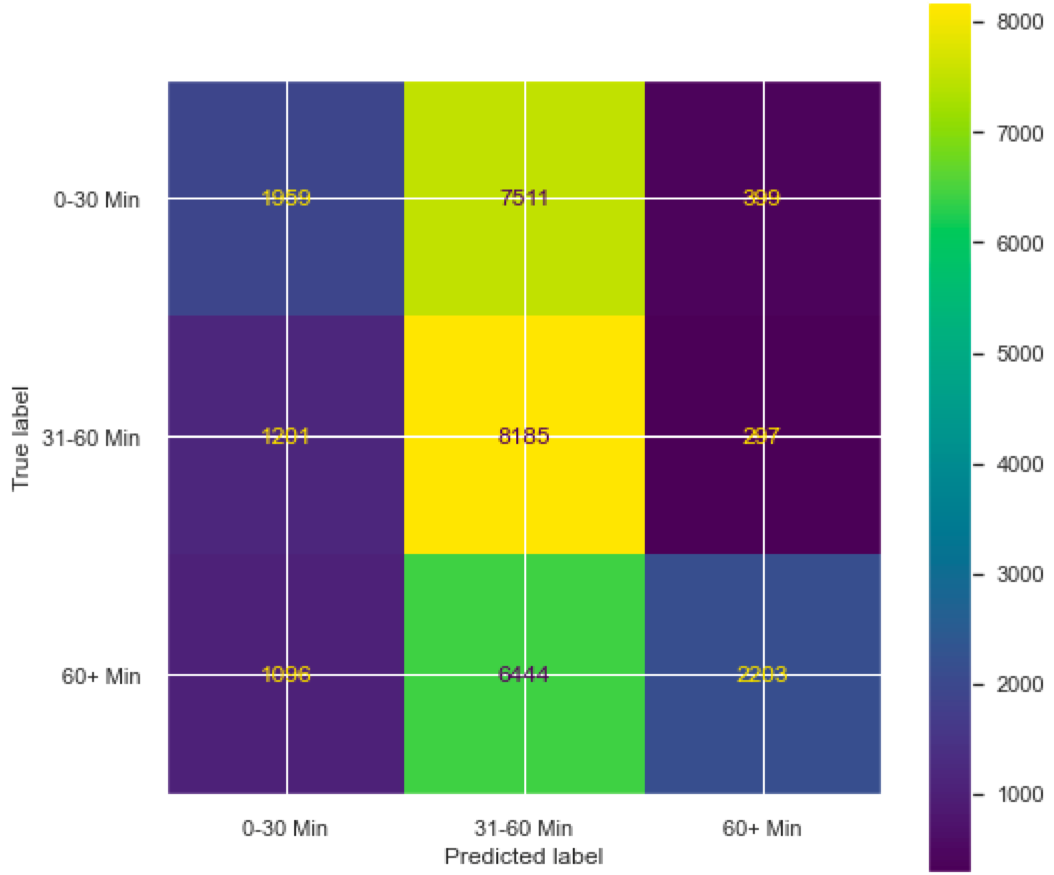


Model 2 - Naive Bayes

After Oversampling

- After oversampling, the NB model gives us a better representation of the data, but the overall scores across the classes are relatively low in comparison to the KNN model. The accuracy here is 0.421 whereas the accuracy in the KNN model was 0.963 (more than double). Similarly the precision, recall and f1 scores are also relatively lower in comparison to the KNN model.
- Conclusion: KNN outperformed the Naive Bayes model in this classification.

	precision	recall	f1-score	support
0-30 Min	0.460	0.199	0.277	9869
31-60 Min	0.370	0.845	0.514	9683
60+ Min	0.760	0.226	0.349	9743
accuracy			0.421	29295
macro avg	0.530	0.423	0.380	29295
weighted avg	0.530	0.421	0.379	29295



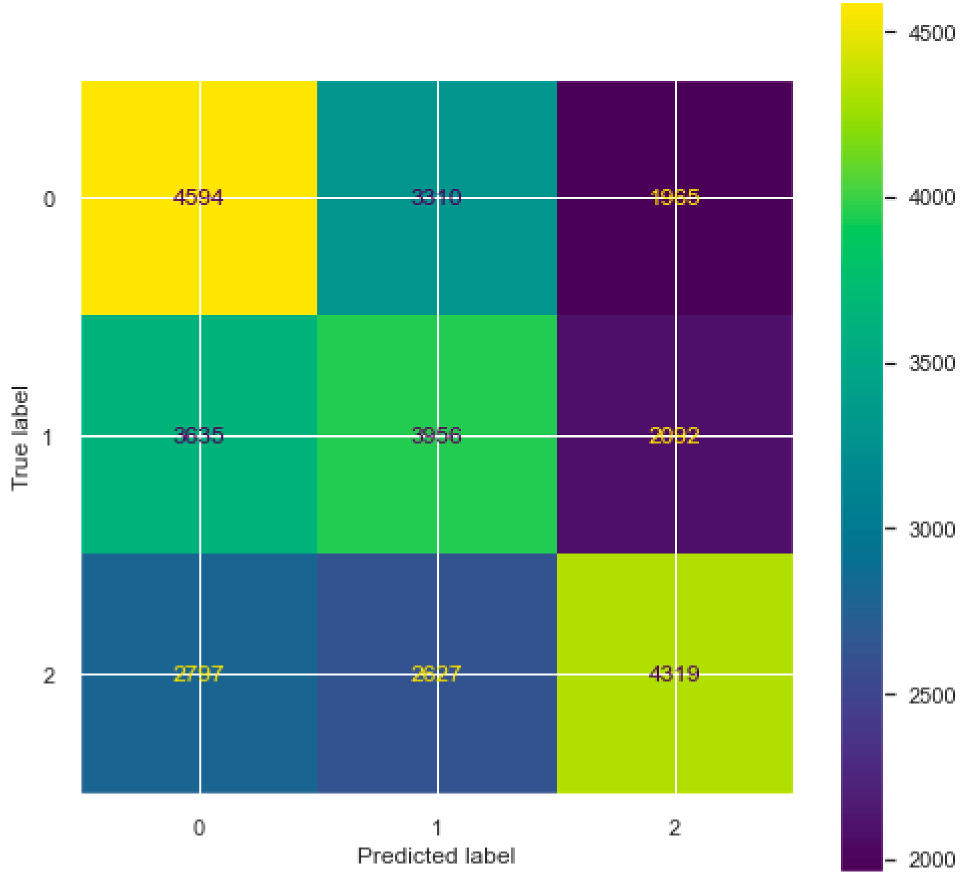
Model 3 - Logistic Regression

After Oversampling

- After oversampling, the logistic regression model gives us a better representation of the data, but the overall scores across the classes are still relatively low in comparison to the KNN model. Similarly the precision, recall and f1 scores are also relatively lower in comparison to the KNN model.

- KNN model outperformed both logistics regression and Naive Bayes Model

	precision	recall	f1-score	support
0-30 Min	0.417	0.465	0.440	9869
31-60 Min	0.400	0.409	0.404	9683
60+ Min	0.516	0.443	0.477	9743
accuracy			0.439	29295
macro avg	0.444	0.439	0.440	29295
weighted avg	0.444	0.439	0.440	29295



Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Question 2 - Classification for Frisk Occurrence

PCA was applied to these models given the large number of features we have in the dataset. This technique could potentially provide better results into out models.

Summary of models used:

- KNN without PCA
- KNN with PCA
- Naive Bayes without PCA
- Naive Bayes with PCA
- Logistic Regression without PCA
- Logistic Regression with PCA

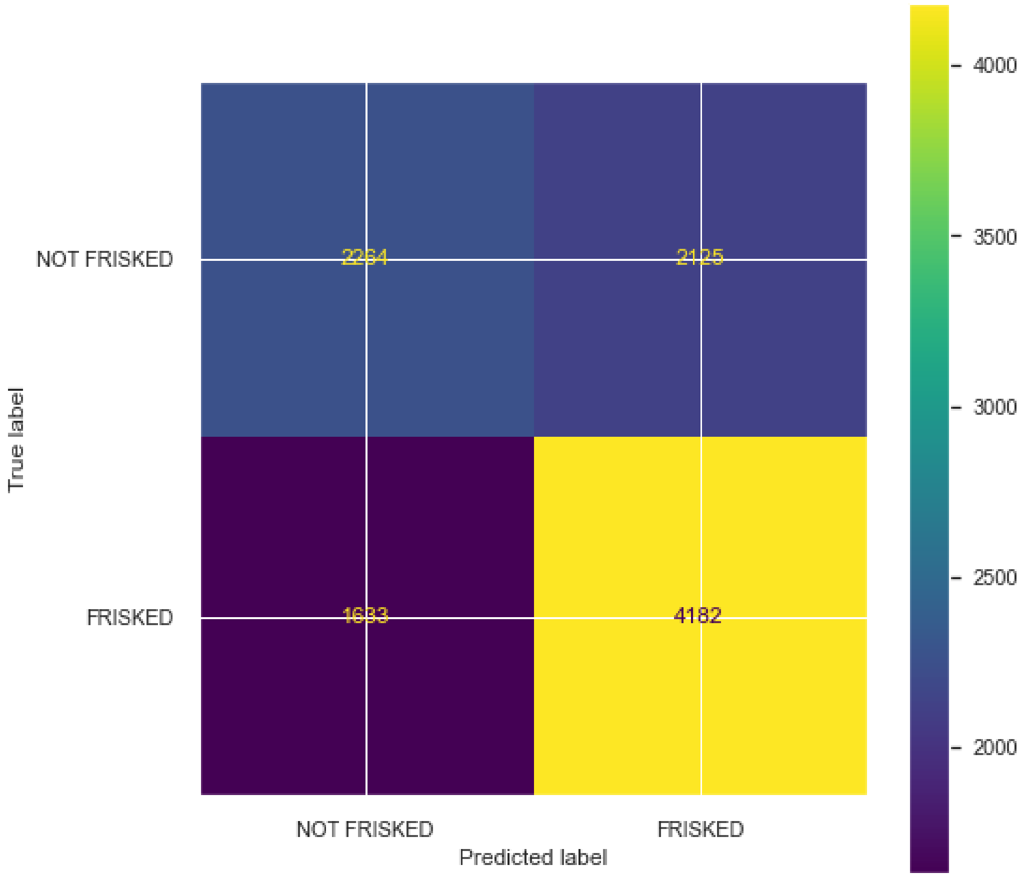
Model 1 - KNN

After PCA

1. Upon using PCA for feature extraction and rerunning the KNN model, the results have definitely become better across the board.

1. There is a significant improvement in the accuracy(increased from 0.63 to 0.712), precision, recall and f1 scores across the classes.

	precision	recall	f1-score	support
0	0.658	0.686	0.672	4389
1	0.755	0.731	0.743	5815
accuracy			0.712	10204
macro avg	0.707	0.709	0.707	10204
weighted avg	0.713	0.712	0.712	10204

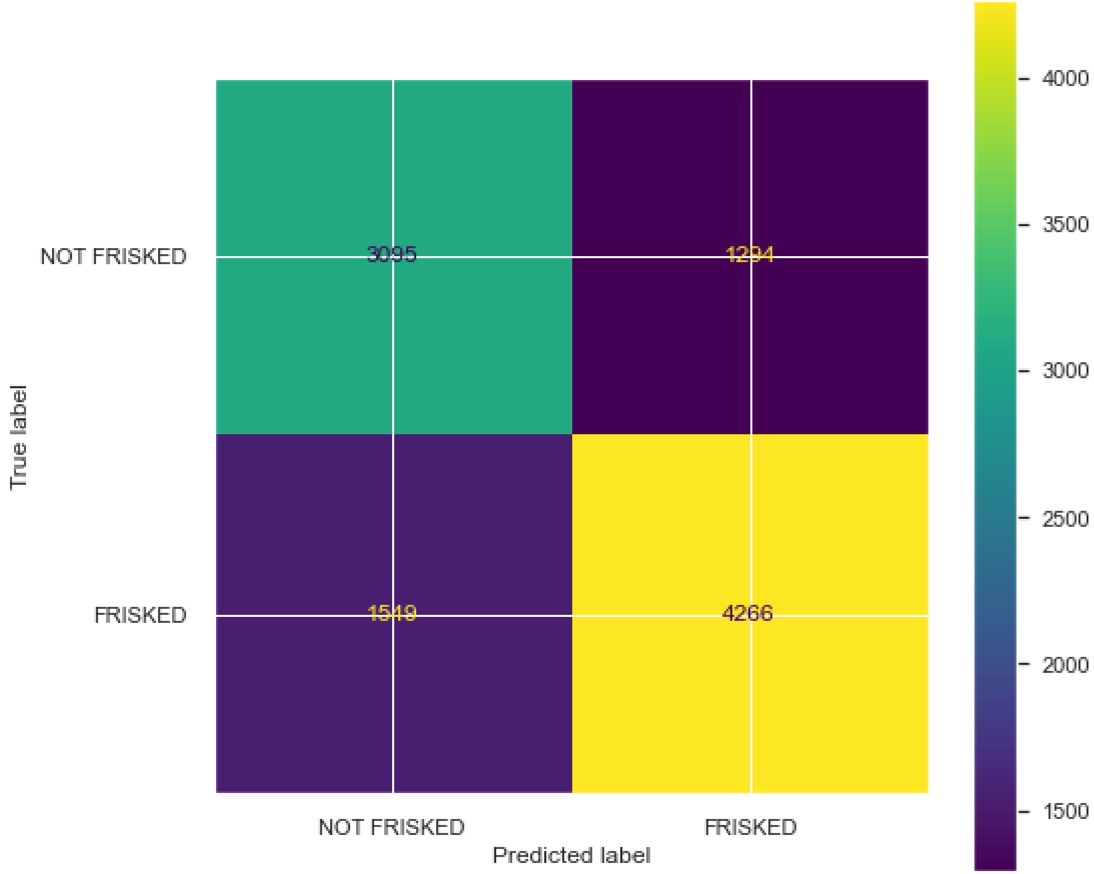


Model 1 - KNN

After Refitting with Best KNN

1. We can conclude that KNN with Principal Component Analysis produced the best results and the accuracy score increased to 72% from 63%.
2. The f1 score for not frisked and frisked increased to 69% and 75% respectively as well, signifying an overall better performance

	precision	recall	f1-score	support
0	0.666	0.705	0.685	4389
1	0.767	0.734	0.750	5815
accuracy			0.721	10204
macro avg	0.717	0.719	0.718	10204
weighted avg	0.724	0.721	0.722	10204



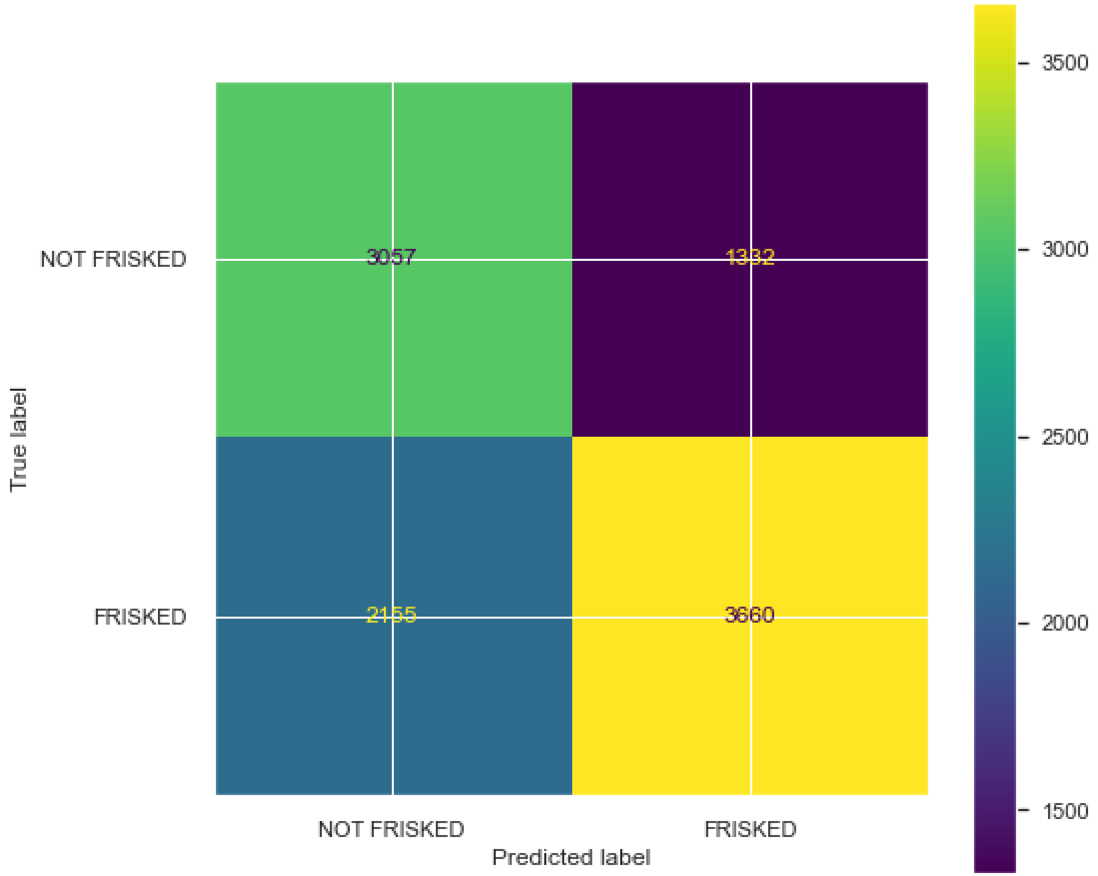
Model 2 - Naive Bayes

Before PCA

1. Similar to KNN, the NB model does a decent job in terms of accuracy and f1 score, while there is definitely a room for improvement.

1. However upon conducting PCA, we realized the the NB model performed better without using PCA for feature extraction.

	precision	recall	f1-score	support
0	0.587	0.697	0.637	4389
1	0.733	0.629	0.677	5815
accuracy			0.658	10204
macro avg	0.660	0.663	0.657	10204
weighted avg	0.670	0.658	0.660	10204

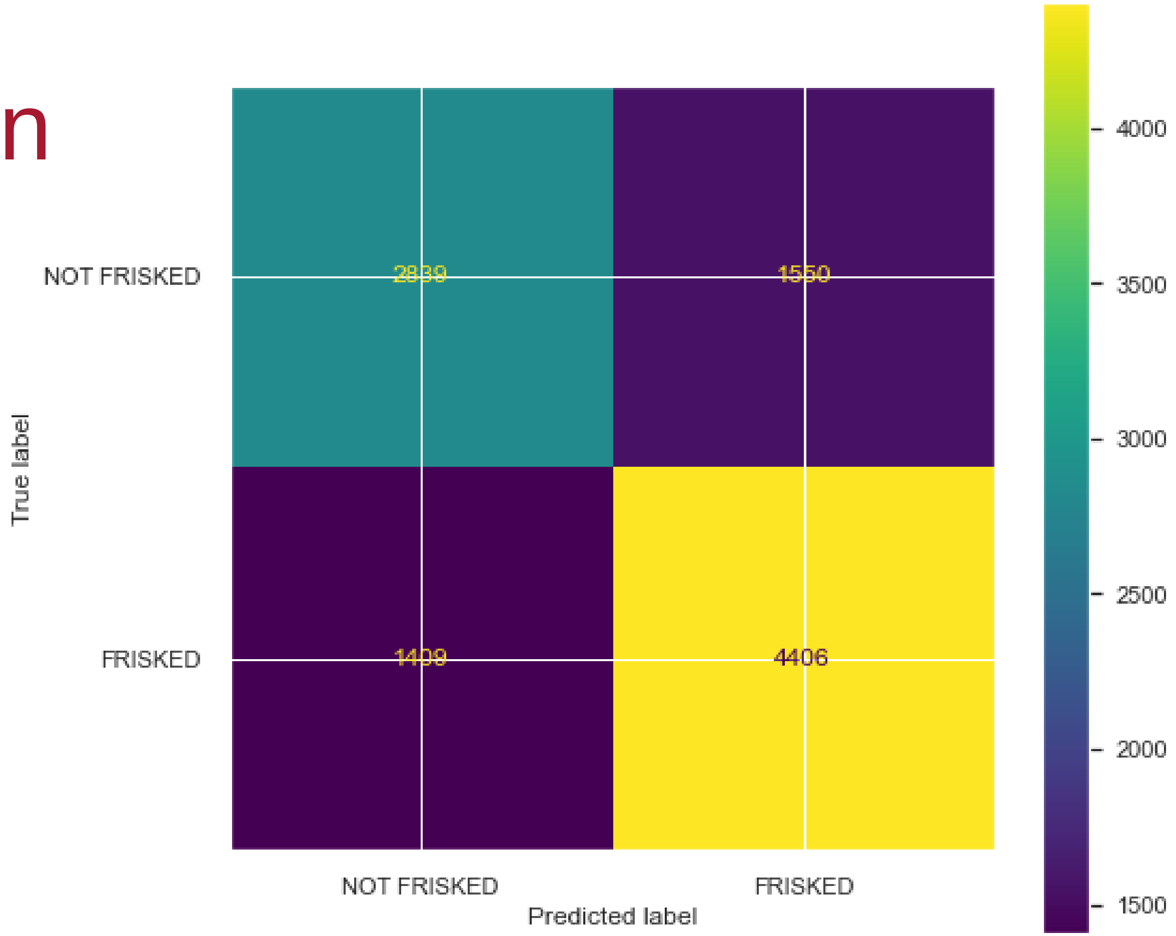


Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Model 3 - Logistic Regression

After PCA

In logistic regression model, the results after conducting the PCA was better than the results before conducting PCA.



	precision	recall	f1-score	support
0	0.668	0.647	0.657	4389
1	0.740	0.758	0.749	5815
accuracy			0.710	10204
macro avg	0.704	0.702	0.703	10204
weighted avg	0.709	0.710	0.709	10204

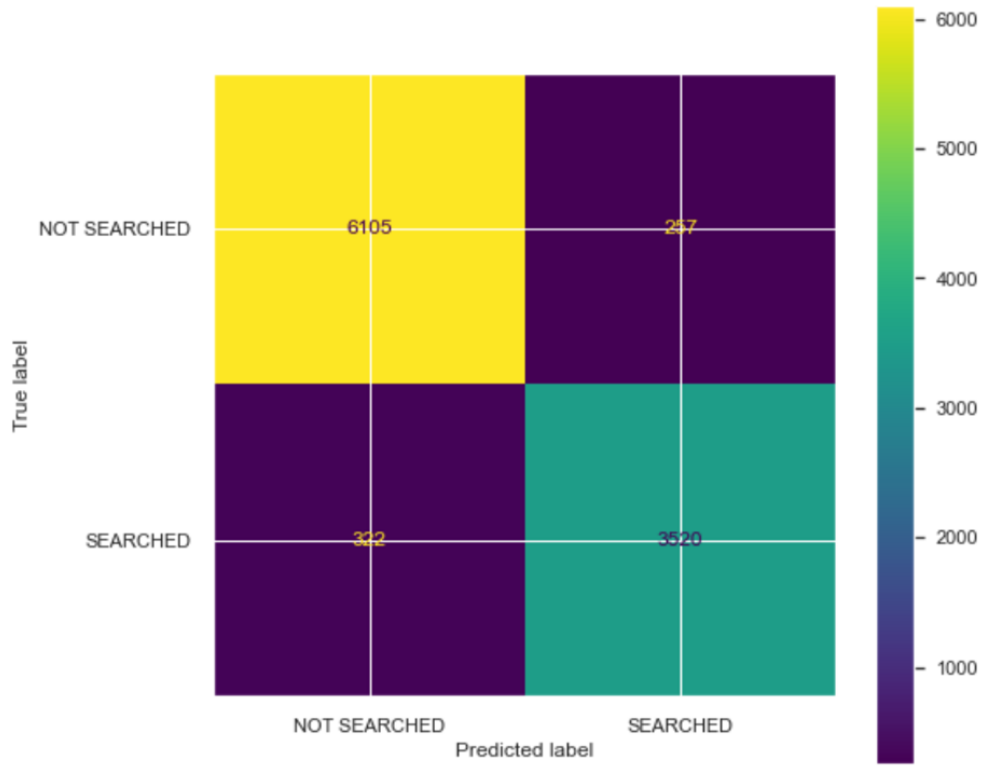
Question 3 - Classification of Searched Flag

PCA was applied into these models given the large number of features we have in the dataset. This technique could potentially provide better results into out models.

Summary of models used:

- KNN without PCA
- KNN with PCA
- Naive Bayes without PCA
- Naive Bayes with PCA
- Logistic Regression without PCA
- Logistic Regression with PCA

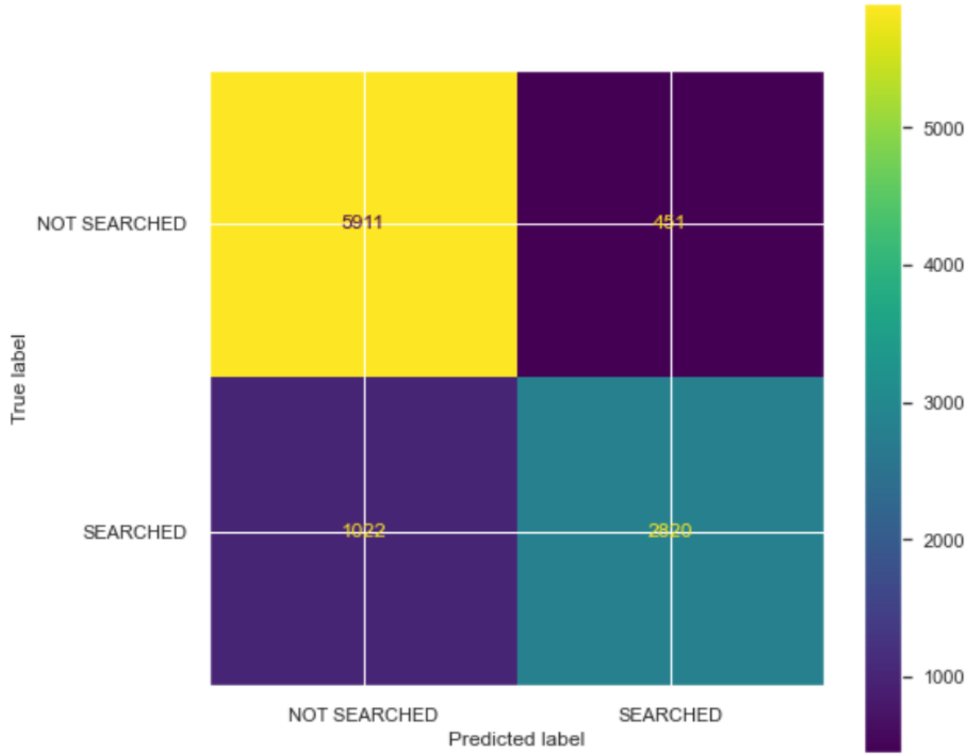
Model 1 - KNN with PCA - Best KNN (3)



	precision	recall	f1-score	support
0	0.950	0.960	0.955	6362
1	0.932	0.916	0.924	3842
accuracy			0.943	10204
macro avg	0.941	0.938	0.939	10204
weighted avg	0.943	0.943	0.943	10204

- By applying PCA, the accuracy, prediction, recall and f1-score of the classes increased significantly from 60% to above 90%. This was observed both with and without the optional number of neighbors.
- There was not a significant difference in the values before and after the model was tuned with the best number of neighbors.

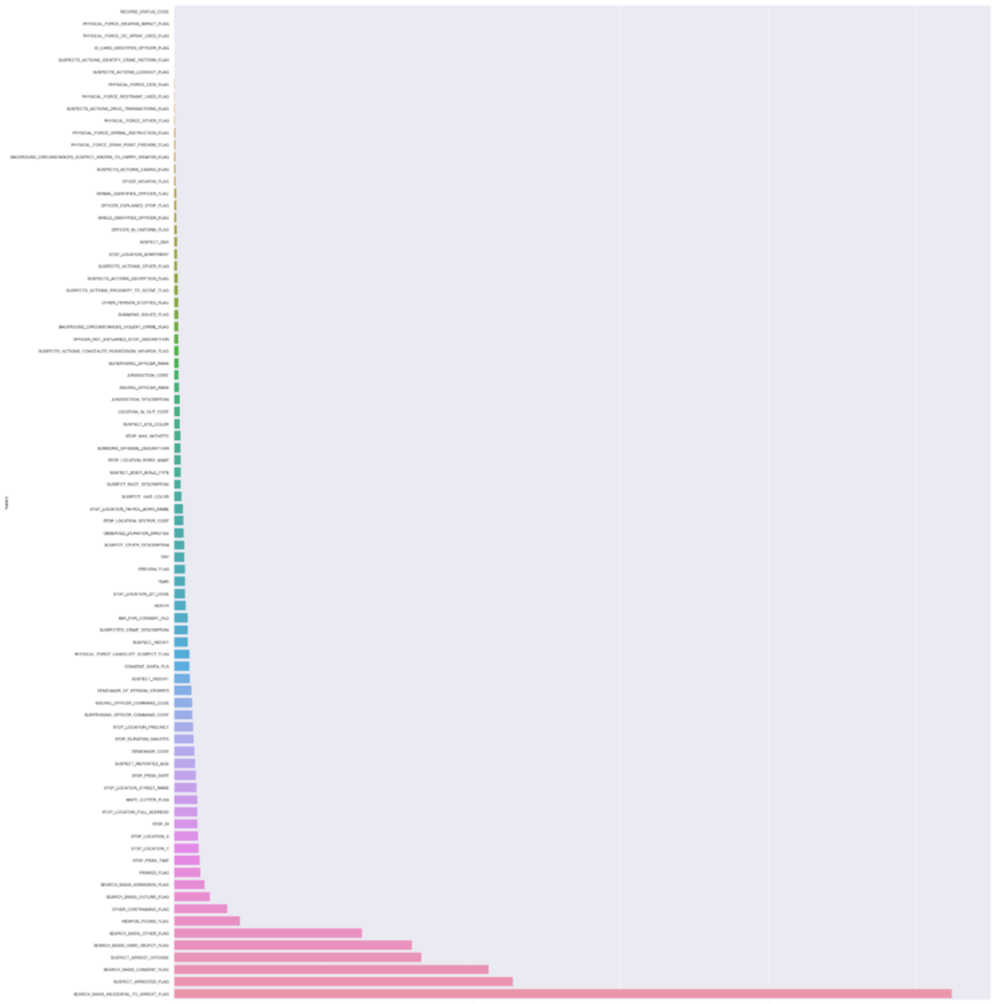
Model 2 - Naive Bayes with PCA



	precision	recall	f1-score	support
0	0.853	0.929	0.889	6362
1	0.862	0.734	0.793	3842
accuracy			0.856	10204
macro avg	0.857	0.832	0.841	10204
weighted avg	0.856	0.856	0.853	10204

- Accuracy scores and precision scores for both the search and not searched classes increased with PCA from 72% to values around 85% and 86%.
- Recall and F1-Score improved significantly for those who were searched - from 42% and 53% to 73% and 79% respectively

Model 3 - Trees - RF with best estimator (160)



- Not significant changes from other Decision Trees and RF ran
- Among all of the models we have run up to this point, this one is providing the best results in terms of accuracy and precision to classify our model.
- The Top 4 most relevant features to classify the searched flag are:
  - SEARCH_BASIS_INCIDDENTAL_TO_ARREST_FLAG
  - SUSPECT_ARRESTED_FLAG,
  - SEARCH_BASIS_CONSENT_FLAG
  - SUSPECT_ARREST_FLAG

	precision	recall	f1-score	support
0	0.999	0.975	0.987	6362
1	0.959	0.999	0.979	3842
accuracy			0.984	10204
macro avg	0.979	0.987	0.983	10204
weighted avg	0.984	0.984	0.984	10204

CONCLUSIONS

Classifying Stop Duration Minutes

1. While the logistic regression model does give a better performance than the Naive Bayes model, our best classification model for the Stop Duration Minutes is the KNN Model(with oversampling).
2. The results have been displayed below for convenient review for the best results obtained for each model type:

i. ***KNN Model(with oversampling):***

	precision	recall	f1-score	support
0-30 Min	1.000	0.891	0.942	9869
31-60 Min	0.917	1.000	0.957	9683
60+ Min	0.980	1.000	0.990	9743
accuracy			0.963	29295
macro avg	0.966	0.964	0.963	29295
weighted avg	0.966	0.963	0.963	29295

BEST

ii. *Naive Bayes model (with oversampling):*

	precision	recall	f1-score	support
0-30 Min	0.460	0.199	0.277	9869
31-60 Min	0.370	0.845	0.514	9683
60+ Min	0.760	0.226	0.349	9743
accuracy			0.421	29295
macro avg	0.530	0.423	0.380	29295
weighted avg	0.530	0.421	0.379	29295

iii. *Logistic Regression model (with oversampling)*

	precision	recall	f1-score	support
0-30 Min	0.417	0.465	0.440	9869
31-60 Min	0.400	0.409	0.404	9683
60+ Min	0.516	0.443	0.477	9743
accuracy			0.439	29295
macro avg	0.444	0.439	0.440	29295
weighted avg	0.444	0.439	0.440	29295

Classifying Frisk Occurrence

1. After running the KNN, NB, and Logistic Regression models, we can conclude that the Logistic Regression model with PCA and the KNN model with PCA and CV selection for accuracy score did the best job classifying our data.Below we are including also the matrixes for the best models run for each KNN, NB and Logistic Regression.
2. The results have been displayed below for convenient review for the best results obtained for each model type:

i. ***KNN Model with PCA:***

	precision	recall	f1-score	support
0	0.666	0.705	0.685	4389
1	0.767	0.734	0.750	5815
accuracy			0.721	10204
macro avg	0.717	0.719	0.718	10204
weighted avg	0.724	0.721	0.722	10204

BEST

ii. *Naive Bayes Model with PCA:*

	precision	recall	f1-score	support
0	0.574	0.739	0.646	4389
1	0.749	0.586	0.657	5815
accuracy			0.652	10204
macro avg	0.661	0.663	0.652	10204
weighted avg	0.674	0.652	0.653	10204

iii. ***Logistic Regression with PCA:***

	precision	recall	f1-score	support
0	0.668	0.647	0.657	4389
1	0.740	0.758	0.749	5815
accuracy			0.710	10204
macro avg	0.704	0.702	0.703	10204
weighted avg	0.709	0.710	0.709	10204

BEST

Classifying Search Occurrence

1. After running the KNN, NB, and Decision Trees/Random Forest models, we can conclude that the Random Forest Model with no PCA tuned using the best number of trees as 160 did the best in classifying our data.
2. The results have been displayed below for convenient review for the best results obtained for each model type:

i. *KNN Model with PCA*

	precision	recall	f1-score	support
0	0.950	0.960	0.955	6362
1	0.932	0.916	0.924	3842
accuracy			0.943	10204
macro avg	0.941	0.938	0.939	10204
weighted avg	0.943	0.943	0.943	10204

iii. * **Random Forest with best number of estimators (160):***

	precision	recall	f1-score	support
0	0.999	0.975	0.987	6362
1	0.959	0.999	0.979	3842
accuracy			0.984	10204
macro avg	0.979	0.987	0.983	10204
weighted avg	0.984	0.984	0.984	10204

BEST

ii. *Naive Bayes Model with PCA:*

	precision	recall	f1-score	support
0	0.853	0.929	0.889	6362
1	0.862	0.734	0.793	3842
accuracy			0.856	10204
macro avg	0.857	0.832	0.841	10204
weighted avg	0.856	0.856	0.853	10204

Objective Conclusions

1. We can predict that an individual who is stopped would fall into one of the 3 classes of stop duration times with an accuracy of 96.3% with an average precision of 96.6% (using a KNN classifier)

1. We can classify whether an individual who is stopped would also be frisked up to an accuracy of 72.1% with an average precision of 71.7% (Using a KNN classifier). We can also use a Logistic Regression model with an accuracy of up to 71% and an average precision of 70.4%.

1. We can classify whether an individual who is stopped would also be searched with an accuracy of up to 98.4% with an average precision of 99% for those who were not searched, and with a 95.9 to those who were searched. For the classification on the search flag, the most important features were:
   a. SEARCH_BASIS_INCIDDENTAL_TO_ARREST_FLAG, SUSPECT_ARRESTED_FLAG, SEARCH_BASIS_CONSENT_FLAG and SUSPECT_ARREST_FLAG, while SUSPECT_SEX, SUSPECT_RACE and SUSPECT_AGE were among the less importants.
   b. This seems to indicate that when searching an individual, the police is not discriminating agaisnt or targetting a particular minority, gender, or age group

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Future Work

Direct Next Steps - Expanding the scope of our analysis from classification to prediction where:

1. We can predict the factors involved in an individual getting stopped for a specific duration of time, and whether there are certain indicators which would skew the possibility of an individual being stopped for a longer period.
2. We are able to extract the factors involved in an individual who is stopped and frisked, and predict the probability that a person with similar characteristics would be exposed to the same situation in the future
3. We can predict the probability that an individual with a specific set of characteristics would be stopped and searched
4. We can increase the size of our superset and track the data all the way back to 2003, increasing the number of datapoints, and thus being able to train our models better.

Future Work

Future Scope

1. Identifying specific locations or regions which are susceptible to a greater number of people being stopped and frisked, and checking its correlation to the relevant geographical, social and economic factors.
2. Predicting the probability of future arrests in a particular location, based on crime statistics.
3. Extending research into sentiment analysis to develop a relationship between social media trends of hate speech and racial profiling with rising police brutality cases.
4. Developing an automated system to cross verify the validity of a stop and search based on specific features. This could be used to eliminate human bias, and eliminate the chance of racial profiling.

References

- Used prior homework and labs as reference
- https://stackoverflow.com/questions/36226083/
- https://www.geeksforgeeks.org/python-pandas-dataframe-astype/
- https://www.statology.org/principal-components-regression-in-python/
- https://www.geeksforgeeks.org/ways-to-filter-pandas-dataframe-by-column-values/
- http://www.science.smith.edu/~jcrouser/SDS293/labs/lab10-py.html
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html?-highlight=linear%20regression#sklearn.linear_model.LinearRegression
- https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
- https://scikit-learn.org/stable/modules/naive_bayes.html
- https://www.washingtonpost.com/news/wonk/wp/2013/08/13/heres-what-you-need-to-know-about-stop-and-frisk-and-why-the-courts-shut-it-down/
- https://www.nyclu.org/en/stop-and-frisk-data
- https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page
- https://www.icpsr.umich.edu/web/NACJD/studies/21660/summary
- https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
- https://www.kaggle.com/sociopath00/random-forest-using-gridsearchcv
- https://www.geeksforgeeks.org/how-to-randomly-select-rows-of-an-array-in-python-with-numpy/
- https://www.scikit-yb.org/en/latest/quickstart.html
- https://www.scikit-yb.org/en/latest/api/features/rankd.html
- https://www.scikit-yb.org/en/latest/api/features/index.html
- https://www.scikit-yb.org/en/latest/api/cluster/silhouette.html
- https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
- https://select-statistics.co.uk/blog/market-basket-analysis-understanding-customer-behaviour/
- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
- https://imbalanced-learn.org/stable/

Video Recording

Zoom video:

In order to access our presentation recording, please click on the following link:

https://cmu.zoom.us/rec/share/0omcziGVvTG_eb7Qi57VrfOzhofO4MlMxcEOjARLoBlaRfLFYpzW2O5ELzHpabbc.kOwHf2T8Qsp7NXXA?startTime=1639098947000

Use the following password to access it: NYS&F2020

If this does not work, please access using through this **Google drive link**:

https://drive.google.com/file/d/1hWrcjzgl-A0v6WhvDEfs0Wq-JsiuYUfv/vie